AD_____

Award Number: W81XWH-11-2-0133

TITLE: Framework for Smart Electronic Health Record-Linked Predictive Models to Optimize Care for Complex Digestive Diseases

PRINCIPAL INVESTIGATOR: Michael A. Dunn, MD

CONTRACTING ORGANIZATION: The University of Pittsburgh

Pittsburgh, PA 15213-3320

REPORT DATE: June 2013

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
June 2013	Annual	12 May 2012 – 11 May 2013

4. TITLE AND SUBTITLE

Framework for Smart Electronic Health Record-Linked Predictive Models to Optimize Care for Complex Digestive Diseases

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-2-0133

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Michael Dunn, MD, Melissa Saul, MS

E-Mail: dunnma@upmc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Pittsburgh
Pittsburgh, PA 15213-3320

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Our major objective is to develop an electronic application capable of integrating and semantically standardizing electronic medical record (EMR) data to generate de-identified datasets populated with longitudinal clinical data drawn from diverse sources. In Year 1 of our project, we have successfully built the infrastructure to support this project. We have defined and generated the EMR-based datasets to be used for algorithm development.
In year 2, we used the EMR output and selected genetic information to construct predictive models of the outcomes of complex digestive diseases using Bayesian network (BN) analysis of the generated databases. We plan on comparing performance among models generated using EMR data alone and data from disease-specific clinical research repositories (with and without genetic data). In collaboration with Walter Reed National Military Medical Center, we will share our data acquisition strategies and algorithmic model development. The integration of the two distinct patient populations will lay the groundwork for future data-sharing projects of mutual interest.

15. SUBJECT TERMS none provided

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	12	19b. TELEPHONE NUMBER (include area code)

Contents

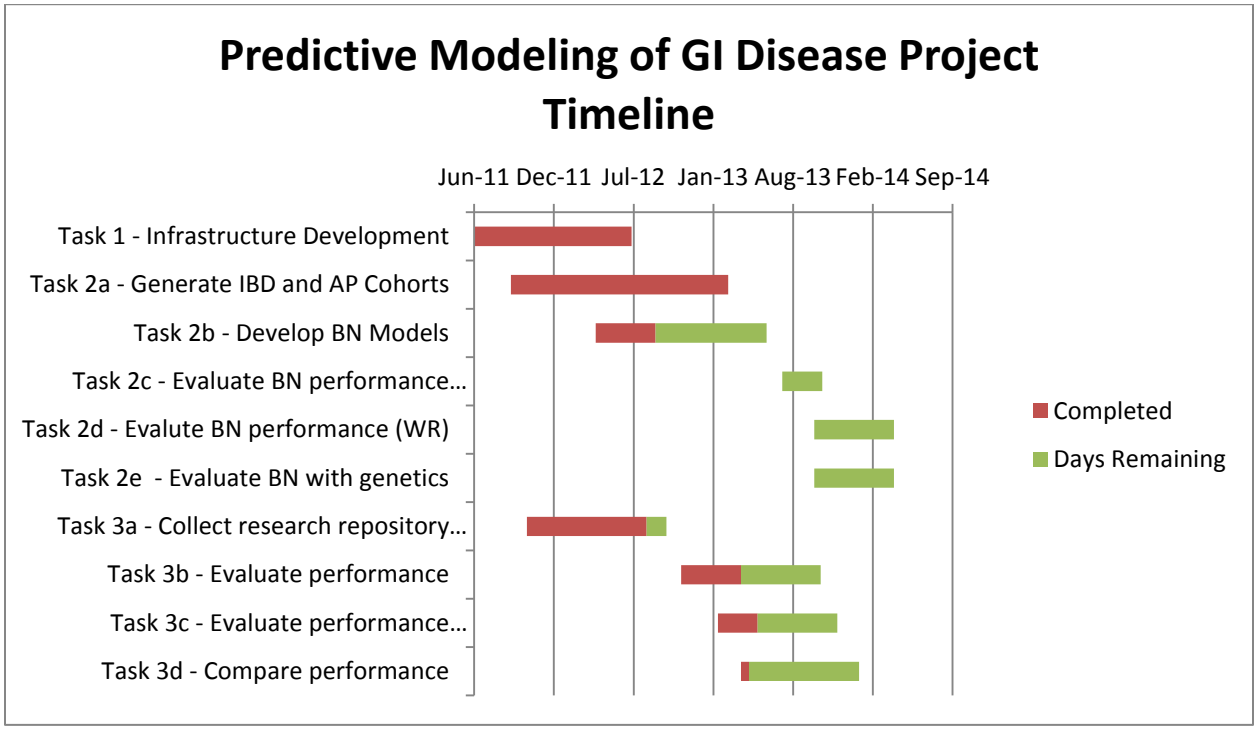
Introduction

Complex disorders result from the interaction of genetic, metabolic, and environmental factors that may not by themselves produce disease but can combine to alter disease severity and its progression. These factors, which may be contained in an electronic medical record (EMR) system, can be used to build predictive models of disease with the hope of improving disease management.

It is difficult to find these factors in EMR systems as the information is in both structured and unstructured formats that have been collected over many years. Research studies, in contrast, only collect a limited snapshot of a patient's clinical history. This information is usually not rich enough to develop predictive models. To construct a useful patient profile for analysis requires collecting disease progression and treatment information from a wide variety of sources that may span twenty years or more.

Our study goal is to develop the Megascope application to provide a software platform for the integration of clinical, genomic and research data collected from multiple sources. The University of Pittsburgh's Department of Biomedical Informatics (DBMI) and Division of Gastroenterology is an ideal collaboration to achieve this goal given our history of successful development of informatics applications and clinical research in complex GI diseases.

We will test the ability of Megascope to support predictive modeling of the outcomes of complex digestive diseases using Bayesian network (BN) analysis of the generated databases. We will further compare performance among models generated using EMR data alone and data from disease-specific clinical research repositories (with and without genetic data).



Body

Progress report on Technical Objective 1 – Infrastructure Development

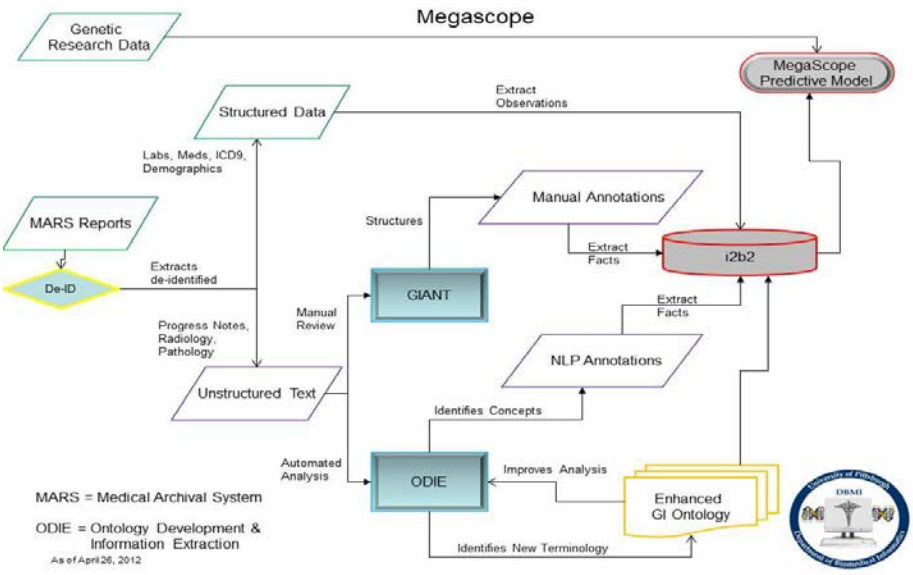
In our initial grant award year, we successfully built an application called Megascope to support our project goals. We realized early in the project that we needed to have a robust, open-source platform that would support the integration of clinical and genetic data. We also wanted to have an application that would not require a lengthy development cycle for creating data models. We decided to use the i2b2 (www.i2b2.org) framework to aggregate various sources of clinical and genomic data into a common vocabulary. This conversion to a common vocabulary, technically referred to as a controlled vocabulary or ontology, allows for us to treat many sources of data as though they are one. The i2b2 structure also has support for data mappings using LOINC and RxNORM enabling the data to be stored in uniform nomenclature.

The i2b2 data model is based on the "star schema" design where each row in the main database table represents a single "fact". The facts are observations about a patient. Observations about a patient are recorded regarding a specific concept such as a lab value or medication order in the context of either an inpatient or outpatient encounter. This way of expressing a concept as an attribute in a row is known as the entity-attribute-value (EAV) model. It is very efficient to query data arranged in a star schema represented in an EAV format as a single index enables all patients' data to be searched in one query. A screen shot of our i2b2 instance is listed in Figure 1.

The i2b2 platform is used by the NIH Clinical Translational Science Award (CTSA) network, and other academic health centers. i2b2 is funded as a cooperative agreement with the National Institutes of Health. The i2b2 platform will 1) enable data sharing across institutions; 2) construct extensible frameworks; 3) be able to utilize existing client and web interfaces and 4) make use of a controlled vocabulary.

To support the natural language processing needed for our analysis, we built our GI clinical phenotyping pipeline (figure 3) by modifying our existing components to develop an information extraction pipeline specific to GI phenotyping. A central component of our pipeline uses our previous work developing the Ontology Development and Information Extraction (ODIE) system for ontology based annotation of clinical documents. The ODIE toolkit encompasses a suite of services for ontology-based text annotation (OA) and ontology enrichment (OE) combined with the ODIE workbench for user interaction, analysis, and visualization. Analysis engines for OA and OE, are executed in the Unstructured Information Management Architecture (UIMA) environment, an open-source, Apache-supported component software platform for unstructured information analysis.

To date, our i2b2 system contains laboratory, demographics, pathology, medication (prescription) data and ICD9 diagnoses and procedure codes and annotated data. We need to add the genetic information after we determine its usefulness in the model. The Megascope application is displayed below:



GIANT

As part of the GI clinical phenotyping pipeline to capture those data elements that are only identifiable by domain experts, we developed a web-based annotation tool, GIANT, to enable researchers to annotate de-identified clinical reports. The application design focuses on providing users with an intelligent workspace, by displaying annotation forms and de-identified reports with the same view, automatic report queuing and providing easy access to annotation guidelines and data definitions. The application produces user statistics to report agreement between multiple annotators who are reviewing the same report. Our tool was built using the Django (www.djangoproject.com) web framework, which is an open-source project built on the Python (www.python.org) programming language. The annotation tool features include controlled user access, database support, progress reporting, task-specific error checking and a site administration interface.

There are two output streams for GIANT. The first output is the report annotations completed by the clinical expert that will be imported into i2b2. The second output is the list of concepts identified in ODIE that appear most frequently in documents. This concept generator is used for feature selection to comprise the elements in the predictive model.

GIANT is currently supporting 16 projects throughout the health system including 5 additional projects in the Division of Gastroenterology and 11 projects in the departments of Pharmacy and Therapeutics, Radiology, and Medicine.

Ontology development and enrichment

As we began examining the operative notes that were annotated in GIANT for Crohn's disease surgery, we recognized that the operative procedure names were complex. In processing the notes through ODIE, we could not find an ontology which recognized some of the procedure names. In discussing this issue with ontology domain experts, we realized that the GI surgery domain is not well represented in standard ontologies. So, we are adding each of the procedure terms to our ontology and will contribute this ontology to the National Center for Biomedical Ontology (www.bioontology.org) upon completion. The same condition exists with identifying acute pancreatitis (AP) in radiology reports.

ODIE identifies both concepts (CUI) and semantic types (TUI) found in the narrative reports. These data will be used as the input for Technical Objective 2 (below).

Progress Report on Technical Objective 2: Algorithm Development

In order to create variables needed for the algorithm development in both Crohn's Disease and AP cohorts, we are utilizing our phenotyping pipeline to classify concepts to specific outcomes and disease severities.

IBD Cohort: We identified the specific outcome (surgery) for our Crohn's set by processing the operative reports through our phenotyping pipeline. The classification system we built for this task was the focus of the manuscript submitted to the Journal of the American Medical Informatics Association (JAMIA) in April, 2013. During our meeting with our Walter Reed colleagues, we discovered that the operative notes are not available in the Army cohort and we needed to modify our approach to identifying surgical events. Our revised approach is to identify a surgical outcome via a surgical pathology report since pathology reports are available on the military cohort. We plan on analyzing the pathology report data on the Pittsburgh cohort to see if we can obtain comparable results of positive operative note for surgery to positive pathology report for surgery. After the appropriate paperwork has been completed, we will use our phenotyping pipeline on the military surgical pathology reports and provide the output to our Walter Reed colleagues.

AP Cohort: Initially, we identified 5970 inpatient visits for 4732 unique patients seen at our institution from 2000 to 2009 who had a primary discharge diagnosis of acute pancreatitis (AP). This was our initial study cohort. However, we did not have as rich clinical data for this cohort as required so we extended our study period to 12/31/2012. This enabled us to add an additional 1770 unique patients to the cohort.

For the AP set, we used our phenotyping pipeline to extract the vital sign data needed to construct the SIRS score along with laboratory and demographic data. The SIRS was constructed on Days 1 and 2 of the hospital stay. At our meeting with the Walter Reed team in April, 2013, we finalized the variables to be used in the AP model:

Variable	**AP Patients (n =)**
Age (mean +/- sd)	
Males - n (%)	
Whites – n (%)	
Charlson co-morbidity Index (median, IQR)	
Etiology – n (%) Biliary Alcohol Both alcohol and biliary Idiopathic Post-ERCP Others	
ICU admission – n (%)	
ICU admission >48 hours – n (%)	
Organ failure – n (%)	
Persistent organ failure – n (%)	
Length of stay – (days) – median (IQR)	
Death (within days of admission) – n (%) 7 days 30 days 90 days	
SIRS Day 1 – n (%)	
SIRS within 48 hours – n (%)	
SIRS Day 2-5 – n (%)	
Persistent SIRS (i.e. >48 hours) – n (%)	

Progress Report on Technical Objective 3: Proof-of-Principle Study

We have completed an electronic review of the 594 patients who are in our NIDDK study. The electronic review was done via GIANT. We identified the patients in this cohort who also have genetic data available in our Immunochip data set. The Immunochip set represents 163 genomic regions of single nucleotide polymorphisms (SNPs) or genetic variations with at least suggestive evidence for association with either

Crohn's disease, ulcerative colitis or both forms of IBD. We analyzed the data using Plink (http://pngu.mgh.harvard.edu/purcell/plink/).

Our IBD cohort is defined in three separate but overlapping groups. There are 1518 individuals who had at least one surgery for Crohn's Disease. Of these 1518, 262 have at least five years of follow-up and had their disease diagnosed within ten years of first being seen at our facility and have genetic information available. We consider this to be our gold set. An additional 339 have genetic information available and at least five years of follow-up but their disease was diagnosed outside of the ten-year window for initial diagnoses. The remaining patients ((n=1017) will be also be used but their analysis may be limited.

Using the 339 patients in the NIDDK cohort, we defined the outcome as one of three choices: 1) no surgery; 2) 1 surgery within 10 years of diagnosis and 3) 2 or more surgeries within 10 years of diagnosis.

Our NIDDK dataset contains 2 variables for a surgical outcome – 1) abdominal surgery captured at the time of enrollment and 2) abdominal surgery that we were able to determine using GIANT (our EHR annotation tool). The addition of the EHR data enabled us to more accurately record surgical outcomes.

For the 339 patients, we obtained genetic information for 332 of them. We have 139 SNPs. Outcomes on 332 individuals (199 with surgical complications at 10 years, 133 without surgical complications at 10 years). There were 154 males and 178 females in the dataset.

In the first cut of the genetic data, we did not find a single genetic variable that was predictive of the surgical outcome. We combined the outcome into a logical variable (having surgery or not having surgery) and redid the analysis. This gave a Bayesian network with 3 genetic variables which has an accuracy of 60% and the area under the ROC curve is 0.648. This level of predictive performance is not great but at least it shows that there is a signal in the genetic variables. Moreover, in most genetic studies of other diseases we have seen so far, the accuracy is in the 60-70% range.

This SNP we discovered is in the mitogen-activated protein kinase kinase kinase 8 gene (MAP3K8). MAP3K8 activates IkappaB kinases, and thus induce the nuclear production of NF-kappaB. MAP3K8 also promotes the production of TNF-alpha and IL-2 during T lymphocyte activation.

Key Research Accomplishments

Abstracts presented at American Gastroenterology Association Digestive Disease Week 2013:

- Visweswaran, S., Saul, M. I., Espino, J. U., Levander, J., Swoger, J. M., Regueiro, M., & Dunn, M. A. (2013). Mo1342 A Concept Recognition Tool to Identify the Surgical Complications of Crohn's Disease in Electronic Health Records. Gastroenterology, 144(5), S-641.
- Yadav, D., Saul, M. I., Papachristou, G. I., Whitcomb, D. C., Visweswaran, S., & Dunn, M. A. (2013). Sa1379 Electronic Health Record (EHR) Information Is Useful to Predict Clinically Relevant Outcomes in Acute Pancreatitis (AP). Gastroenterology, 144(5), S-279.
- Vargas, E.J., Ramos Rivers, C.M., Regueiro, M., Barrie, A., Baidoo, L., Schwartz, M., Swoger, J.L., Dunn, M.A., Dudekula, A., & Binion, D.G. Inflammatory Bowel Disease and Selective Immunoglobulin a Deficiency. Gastroenterology, 144(5) Su1250.
- Umpathy, C., Ramos Rivers, C.M., Regueiro, M., Baidoo, L. Barrie, A., Schwartz, M., Swoger, J.M., Dunn, M.A., Weyant, K.A., Watson, A.R. & Binion, D.G. Effect of the Bile Acid Sequestrant Colesevelam on Health Related Quality of Life in Crohn's Disease. Gastroenterology, 144(5) Su1269.
- Ramos Rivers, C.M., Vargas, E.J., Binion, D.G., Regueiro, M., Baidoo, L., Barrie, A., Swoger, J.M., Schwartz, M., Dunn, M.A., Szigethy, E., & Benhayon, D. Sleep Disturbance and the Clinical Course of IBD. Gastroenterology, 144(5) Mo1274.

- Ramos Rivers, C.M., Eric J. Vargas, E.J., Coates, M., Regueiro, M., Dunn, M.A., Swoger, J. M., Schwartz, M., Barrie, A., Baidoo, L., Szigethy, E., & Binion, D.G. Clinical Factors Contributing to Abdominal Pain in IBD. Gastroenterology, 144(5) Mo1290.
- Baidoo, L., Ramos Rivers, C.M., Regueiro, M, Barrie, A., Swoger, J.M., Schwartz, M., Szigethy, E., Dunn, M.A., Dudekula, A., & Binion, D.G. Improving the Quality of Crohn's Disease Care. Gastroenterology, 144(5) Mo1305.
- Gajendran, M., Weyant, K.A., Ramos Rivers, C.M., Gede, T., Regueiro, M., Baidoo, L., Schwartz, M., Swoger, J.M., Barrie, A., Dunn, M.A., Watson, A.R., & Binion, D.G. Post-Operative Recurrence of Crohn's Disease at the Surgical Anastomosis and Downstream Bowel. Gastroenterology, 144(5) Mo1315.
- Seminerio, J.L., Ramos Rivers, C.M., Weyant, K.A., Regueiro, M., Baidoo, L., Barrie, A., Swoger, J.M., Schwartz, M., Dunn, M.A., & Binion, D.G. Weight Based Dosing in the Obese Inflammatory Bowel Disease Patient. Gastroenterology, 144(5) Mo1344.

Paper submitted to Journal of American Medical Informatics Association:

- Visweswaran, S. Saul, M. Morris M, Espino JU, Levander J, Swoger JM, Regueiro M, Dunn MA. Automated Identification of Complex Phenotypes in Electronic Health Records

Reportable Outcomes

- Coordinated Research meeting for Walter Reed and Pittsburgh co-investigators and staff in Pittsburgh on April 19, 2013 (agenda included in appendix).
- Nine abstracts accepted and presented for Digestive Disorders Week 2013
- One paper submitted to Journal of American Medical Informatics Association special issue on clinical phenotyping
- Initiated paperwork for task order with Kennell and Associates to perform de-identification and text classification on Walter Reed set.

Conclusions

We have assembled a dynamic and strong team from both gastroenterology and informatics. We have built our infrastructure according to plan. We are looking forward to working with the Walter Reed group to test and validate our models.

Figures

Figure 1 i2b2 Query Tool

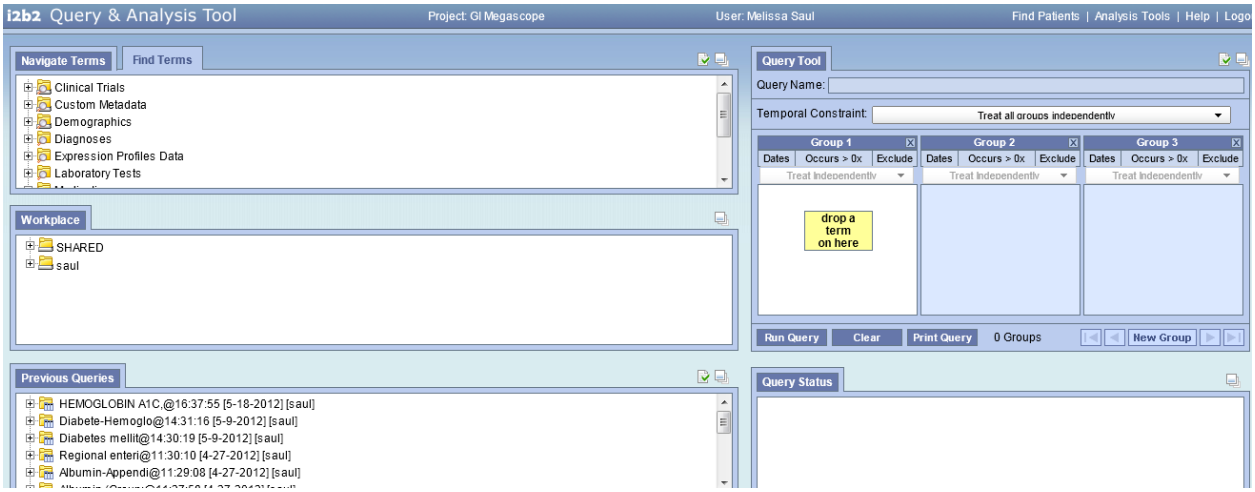


Figure 2 – GIANT user interface

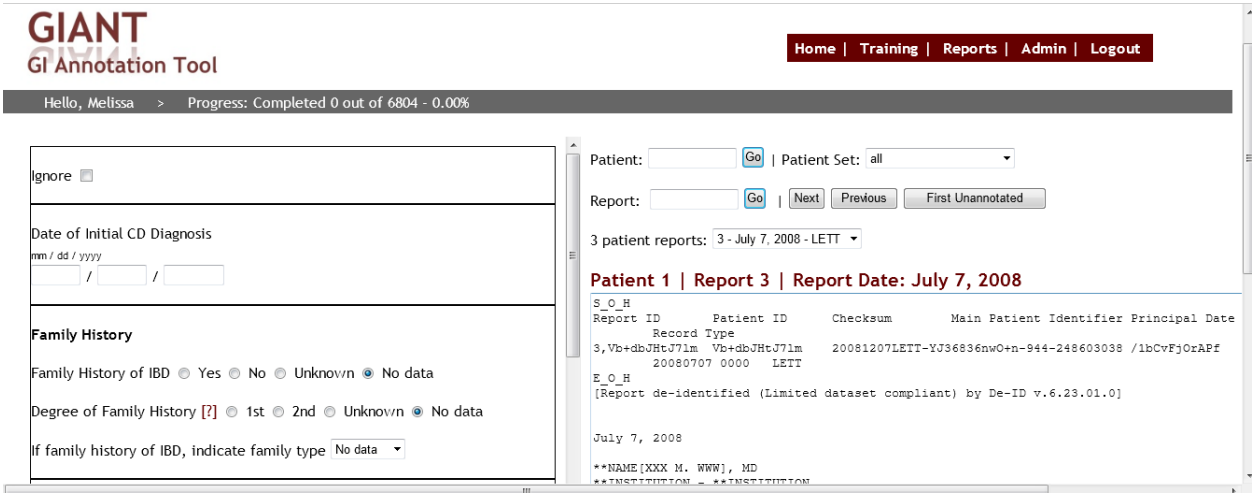
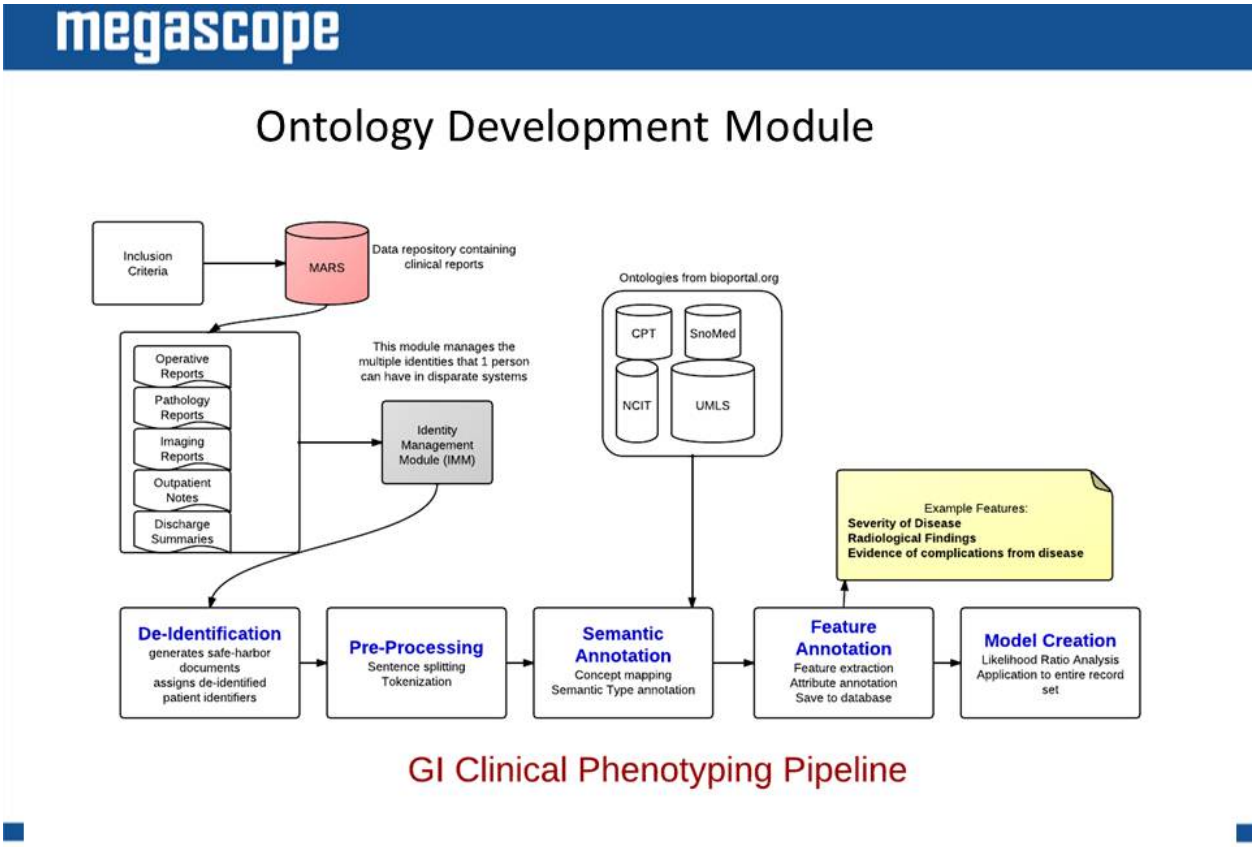


Figure 3 – GI Clinical Phenotyping Pipeline



Appendices

Appendix 1 – Agenda for April 19, 2013 Joint Meeting

Agenda

Framework for Smart Electronic Health Record Linked Predictive Models to Optimize Care for Complex Digestive Diseases Project

Hosted by: University of Pittsburgh Department of Biomedical Informatics (DBMI)
April 19, 2013

I. Status Report of CD and AP Projects – All

II. AP Project Review – Dhiraj Yadav, MD, MPH

III. Bayesian Modeling Methods Workshop – Shyam Visweswaran, MD, PhD
   a. Report on Pittsburgh CD Cohort with genetic data modeling
   b. Hugin software demo

IV. Next Steps

Meeting Information

Location: Department of Biomedical Informatics
5607 Baum Blvd., 5th floor
Pittsburgh, PA 15261

There is a sign indicating DBMI Office Entrance in front of building. The building looks like a construction zone. There is an Aldi's Grocery Store in one corner of the building.

Parking: Parking lot is across the street in abandoned parking lot. Labeled as DBMI Parking

Contact Info

Melissa Saul – 412-818-5448 (cell)